A diffusion limit for a queueing model in the form of a Walsh Brownian motion

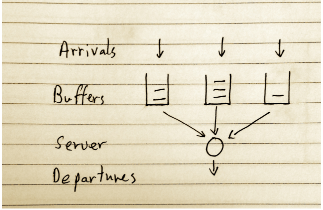
Rami Atar (EE Department, Technion)

Joint work with

Asaf Cohen (Haifa U.)

Banff, 2017

Model



- d buffers, a single server.
- Renewal arrivals with mean interarrival $1/\lambda_i^r$ (finite 2nd moment) for $i \in \{1, 2, \ldots, d\}$, r a scaling parameter.
- For each i, IID job sizes, mean $1/\mu_i^r$ (finite 2nd moment).
- Independence of stochastic primitives
- A *policy* is a rule dictating which job is served at each time.
- *Heavy traffic* asymptotics corresponds to
  (i) a critical load condition,
  (ii) diffusion scale.

Heavy traffic

- Time acceleration

$$\lambda_i^r = \lambda_i r^2 + \hat{\lambda}_i r + o(r)$$
$$\mu_i^r = \mu_i r^2 + \hat{\mu}_i r + o(r).$$

- Critical load

$$\sum_{i=1}^{d} \frac{\lambda_i}{\mu_i} = 1.$$

- Queue length process $Q^r = (Q_1^r, \ldots, Q_d^r)$, well defined once a policy is specified.
- Normalization $\hat{Q}^r = r^{-1} Q^r$.

Some well-understood policies

- *Fixed priority:* buffers ranked and server prioritizes accordingly.
- *Serve the longest queue:* server always selects the longest queue. Motivation: minimize longest delays.
- (One also specifies preemptive or nonpreemptive service and how ties are broken.)

Theorem (Whitt (1971), Reiman (1984))

i. Under fixed priority,

$$(\hat{Q}_1^r, \ldots, \hat{Q}_d^r) \Rightarrow (0, \ldots, 0, R),$$

ii. Under SLQ,

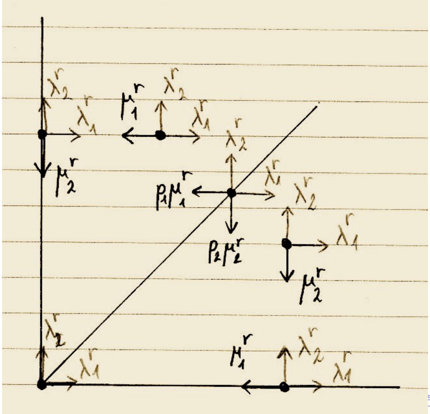
$$(\hat{Q}_1^r, \ldots, \hat{Q}_d^r) \Rightarrow (\tilde{R}, \ldots, \tilde{R}),$$

where R and $\tilde{R}$ are reflected Brownian motion on $[0, \infty)$ (with specific initial condition, drift and diffusion coefficients).

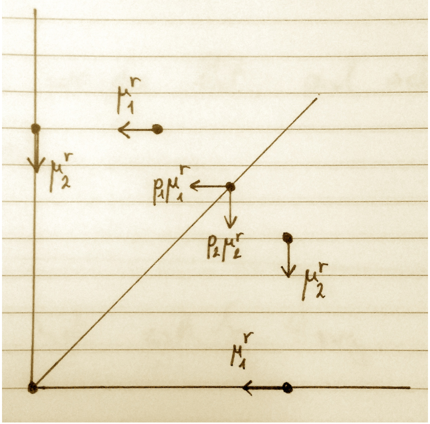
- Laws of R and $\tilde{R}$ determined by first two moments of the primitives.

Serve the shortest queue

- The server always selects the shortest queue. Rationale: minimize the *number* of congested queueing, especially when uncertain about the various traffic intensities.
- Markovian setting (Poisson arrivals, exponential job sizes).
- Tie breaking according to some $\{p_i\}_{i=1,\dots,d}$.



removing the lambdas



Walsh BM

- Proposed by Walsh (1978) as a diffusion process that performs BM (with drift) on a (finite) union of rays emanating from the origin in $\mathbb{R}^2$, in which the entrance law from the origin to the different rays follows a given probability distribution.
- Early results: Rogers (1983), Baxter and Chacon (1984), Varopoulos (1985), Salisbury (1986), Barlow, Pitman and Yor (1989).
- Skew BM: Barlow, Burdzy, Kaspi and Mandelbaum (2000), Burdzy and Chen (2001), Burdzy and Kaspi (2004).
- Recent: Ichiba, Karatsaz, Prokaj and Yan (2015) SDE for Walsh semimartingales.

Walsh BM on S

- Denote $S = \{x \in \mathbb{R}^d_+ : x_i > 0 \text{ for at most one } i\}$.
- Convenient to work with the definition of Barlow, Pitman and Yor (1989) via semigroups. Let R be a (b, σ)-RBM and let q be a probability distribution on $\{1, \ldots, d\}$. Let ζ be the hitting time of R to zero. Then X *is a* (b, σ, q)*-WBM if for* $f \in C_0(S)$ *and* $x = re_{i_0} \in S$,

$$E_x[f(X_t)] = E_r[f(R_t e_{i_0}) 1_{\{t<\zeta\}}] + \sum_i q_i E_0[f(R_t e_i) 1_{\{t \geq \zeta\}}].$$

- Proved by BPY to be a strong Markov, Feller process.

SSQ in heavy traffic

- Define
$$\hat{X}^r = \Big(\frac{\hat{Q}_1^r}{\mu_1}, \dots, \frac{\hat{Q}_d^r}{\mu_d}\Big).$$
- Assume $\hat{X}^r(0)$ converges to a RV supported on S.

Theorem

As $r \to \infty$, $\hat{X}^r \Rightarrow X$, u.o.c., where X is a Walsh BM on S. The modulus $R = 1 \cdot X$ is a RBM with specific (constant) drift and diffusion coefficients.

- (b, σ) are explicit whereas q implicit.
- q expected to depend on data beyond first and second moments.

Literature on the SSQ

Has been proposed for *packet scheduling* on the internet.

- *"Thanks to this simple policy, the scheduler prioritizes constant bit rate flows associated with delay-sensitive applications such as voice and audio/video streaming...; priority is thus implicitly given to smooth flows over data traffic... sending packets in bursts."* Guillemin and Simonian, Orange Labs (2014).
- Has been referred to as 'implicit service differentiation', 'self prioritization of audio and video traffic'.
- Proposed in two ways: queues correspond to different end users, the scheduler is at the base station; queues correspond to different types of data that a single user transmits/receives (scheduler is at the home gateway).
- Experiments show that it performs well (Nasser, Al-Manthari and Hassaneim (2005)).

Mathematical treatment: For $d = 2$ and exponential service times, expressions for the Laplace transform of the stationary distribution (Guillemin and Simonian (2012, 2013)).

Idea of proof

- Convergence toward S: $\sup_{t\in[0,T]} \operatorname{dist}(\hat{X}^r_t, S) \Rightarrow 0$.
  – Reason: see picture.
- $1 \cdot X^r \Rightarrow R$.
  – A standard result (for a general policy).

Remark: C-tightness of $\hat{X}^r$ follows. However the proof does not proceed by analyzing subsequential limits. This is because *strong Markovity* is crucially used. Strong Markovity of WBM cannot be used before establishing that the limit is a WBM; we rely on that of the prelimit.

Lemma

Denote

$$S_i = \{ye_i : y \in \mathbb{R}_+\}$$
$$S_i^\varepsilon = \{x \in \mathbb{R}_+^d : \mathrm{dist}(x, S_i) \le \varepsilon\}$$
$$S^\varepsilon = \{x \in \mathbb{R}_+^d : \mathrm{dist}(x, S) \le \varepsilon\}$$

- $\hat{R}^r(t) = 1 \cdot \hat{X}^r(t)$
- $\tau_\varepsilon^r = \inf\{t : \hat{R}^r(t) \ge \varepsilon\}$

Lemma

There exists $(q_i)_{i=1,\dots,d}$, $1 \cdot q = 1$, *such that*

$$\lim_{\varepsilon \downarrow 0} \limsup_{r\to\infty} |P_0(\hat{X}^r(\tau_\varepsilon^r) \in S_i^\varepsilon) - q_i| = 0.$$

Proof of lemma

First, instead of a double limit it is easier to work with a single one.

- By a change of measure, modify (with little cost) the intensities

$$\lambda_i^r = \lambda_i r^2 + \hat{\lambda}_i r + o(r), \quad \mu_i^r = \mu_i r^2 + \hat{\mu}_i r + o(r)$$

  into

$$\lambda_i^r = \lambda_i r^2, \qquad\qquad \mu_i^r = \mu_i r^2.$$

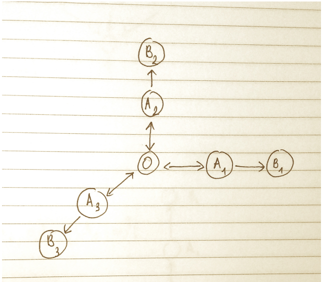
- Then Q^r is a time acceleration of a single process $\hat{Q}$, $Q^r = \hat{Q}(r^2 \cdot)$; $\hat{X}_i = \frac{\hat{Q}_i}{\mu_i}$.
- Let $\tau^r = \inf\{t : 1 \cdot \hat{X}(t) \geq r\}$ and attempt to prove that

$$q_i^r := P_0(\hat{X}(\tau^r) \in S_i^{\varepsilon_0 r})$$

  has a limit.
- It is a Cauchy sequence argument.

A toy model

Consider a Markov chain on $2d + 1$ states. B_i are absorbing.



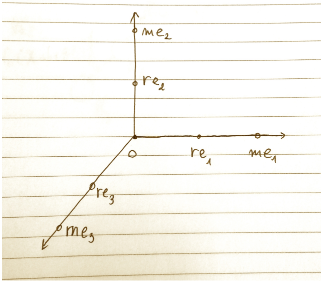
Then

$$\max_i |p(A_i, 0) - p(A_1, 0)|$$

controls

$$\max_i |p(0, A_i) - P_0(\text{getting absorbed at } B_i)|.$$

- Back to the process $\hat{X}$, consider this analogous picture



- Recall $q_i^r = P_0(\hat{X}(\tau^r) \in S_i^{\varepsilon_0 r})$. We aim at showing it is Cauchy via

$$\exists \delta \in (0,1) \text{ s.t. } |q_i^r - q_i^m| \leq \delta^k \text{ for } r \in [2^k, 2^{k+1}],\ m = 2^{k+2},\ \forall k,$$

since this would imply, for general $r < m$,

$$|q_i^r - q_i^m| \leq \sum_{\log_2 r \leq j \leq \log_2 m} 2\delta^j \leq c\delta^{\log_2 r}.$$

- Now make the sleeves r^{1-c} thin, and use the fact that $1 \cdot \hat{X}$ is a martingale to get

$$\forall x \in B(re_i, r^{1-c}), \quad \left| P_x(\zeta < \tau^m) - \frac{m-r}{m} \right| \le r^{-c}.$$

  In view of the toy model this should give estimate that makes $|q_i^r - q_i^m|$ small. However, we need to improve the sleeve estimate from $o(1)$ to r^{-c}, and to obtain similar estimates for the event that the walk switches sleeves without passing through the origin.
- On what time interval are the estimates required? ζ and τ^m (starting in the ball) do not occur within $[0, r^2]$ w.h.p., but within $[0, r^2 \log r]$.
- Hence the estimate we really need is

$$P_x(\|\mathrm{dist}(\hat{X}, S)\|_{r^2 \log r} > r^{-a}) < r^{-c}, \qquad \text{if } \mathrm{dist}(x, S) < \gamma r^{-a}.$$

- This is achieved by working with a suitable Lyapunov function. Measures distance from S and has the intuitive meaning of work present in all but longest queue:

$$F(x) = \sum_i x_i - \max_i x_i.$$

Proof of theorem

One needs to show for $x^r \to x \in S$, uniformly for x in compacts,

$$E_{x^r} f(\hat{X}^r(t)) \to E_x f(X(t)).$$

Focus on $x^r = x = 0$. Fix $\varepsilon > 0$. Let $\zeta_0^r = 0$ and for $m = 0, 1, 2, \ldots$,

$$\tau_m^r = \inf\{t > \zeta_m^r : 1 \cdot X^r(t) \geq \varepsilon\},$$
$$\zeta_{m+1}^r = \inf\{t > \tau_m^r : 1 \cdot X^r(t) = 0\}.$$

$$\begin{aligned}
E_0 f(\hat{X}^r(t)) &\sim \sum_i \sum_m E_0[f(X^r(t)) 1_{\{\tau_m^r \leq t < \zeta_{m+1}^r\}} 1_{\{X^r(\tau_m^r) \in S_i^{r-c}\}}] \\
&\sim \sum_i \sum_m E_0[f(1 \cdot X^r(t)\, e_i) 1_{\{\tau_m^r \leq t < \zeta_{m+1}^r\}} 1_{\{X^r(\tau_m^r) \in S_i^{r-c}\}}] \\
&\sim^* \sum_i \sum_m E_0[f(1 \cdot X^r(t)\, e_i) 1_{\{\tau_m^r \leq t < \zeta_{m+1}^r\}}] q_i \\
&\sim \sum_i E_0[f(R(t) e_i)] q_i
\end{aligned}$$

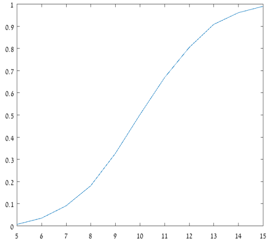
(*) Lemma + another lemma on asymptotic independence, for fixed m, of $X^r(\tau_m^r)$ and τ_m^r.

Two main open questions

- Dependence of the angular distribution q on the data.
- The queueing model is natural to consider for general job size distributions. How to treat it beyond the Markov case?

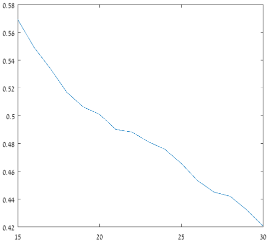
q_1 as a function of λ_1, fixed μ's

$\mu_1 = 20$, $\mu_2 = 20$, $\lambda_1 = 5 \ldots 15$, $\lambda_2 = \mu_2 - \lambda_1$, $p_1 = .5$, $p_2 = .5$



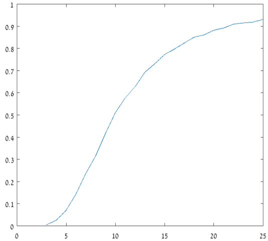
q_1 as a function of μ_1, fixed λ's

$\lambda_1 = 10$, $\lambda_2 = 10$, $\mu_1 = 15 \ldots 30$, $\mu_2 = 1/(1/\lambda_1 - 1/\mu_1)$, $p_1 = .5$, $p_2 = .5$



q_1 as a function of λ_1, fixed λ_1/μ_1, λ_2, μ_2

$\mu_2 = 20$, $\lambda_2 = 10$, $\lambda_1 = 3...25$, $\mu_1 = 2\lambda_1$, $p_1 = .5$, $p_2 = .5$



q_1 as a function of p_1, fixed λ's and μ's

$\lambda_1 = 10$, $\lambda_2 = 10$, $\mu_1 = 20$, $\mu_2 = 20$, $p_1 = 0...1$, $p_2 = 1 - p_1$